Available online at www.sciencedirect.com

ScienceDirect

Energy Procedia 132 (2017) 555–560

Energy Procedia

www.elsevier.com/locate/procedia

11th Nordic Symposium on Building Physics, NSB2017, 11-14 June 2017, Trondheim, Norway

Metamodeling energy indicators in neighborhoods with growing deployment of heat pumps and rooftop photovoltaics

Christina Protopapadaki [a,b,*], Dirk Saelens [a,b]

[a] KU Leuven, Civil Engineering Department, Building Physics Section, Kasteelpark Arenberg 40, 3001 Leuven, Belgium
[b] EnergyVille, Thor Park 8310, 3600 Genk, Belgium

Abstract

This paper evaluates simple metamodels to predict local electricity demand and grid restrictions, in residential neighborhoods with heat pumps and photovoltaics. The procedure and challenges of developing such models are described. Modeling is based on results obtained from detailed simulation of buildings and the grid. Linear and logistic regression models are developed for electricity demand and minimum voltage respectively, as a function of neighborhood characteristics, related to both building and electrical network properties. The paper shows that linear regression can be used for a first evaluation of electricity demand. For voltage violations, logistic regression gives acceptable results; however, more complex models are needed to approximate voltage levels.

Keywords: metamodeling ; linear regression; logistic regression; grid impact; electricity demand; residential neighborhood; heat pump

1. Introduction

Evaluation of building-related energy policy measures, such as refurbishment and renewable energy utilization, is typically performed at building stock level. The challenge in this approach is to account for local technical aspects, such as interactions at the electricity distribution grid, as they can influence the feasibility and effectiveness of policy measures, by limiting, for instance, the permissible penetration rate of heat pumps or distributed generation. Comprehensive dynamic models of building and energy systems at a district level can provide insight into these effects, but are computationally intensive. To overcome the resource and time constraints of detailed simulations, and

* Corresponding author. Tel.: +32 16 3 22283; fax: +32 16 3 21980.
E-mail address: christina.protopapadaki@kuleuven.be

10.1016/j.egypro.2017.09.736

to allow for faster and simpler policy evaluation, surrogate models (*metamodels*) could be employed. This paper, therefore, develops and evaluates basic regression models as metamodels to predict local annual electricity demand and grid voltage violations from detailed simulation results. The response variables are modeled as a function of neighborhood characteristics, related to buildings, as well as electrical network properties for the voltage violations. Data and detailed simulation models for this purpose were available from previous work, for a Belgian residential context, and statistical modeling was performed with the Matlab statistics toolbox. The paper aims to discuss the procedure and challenges of creating such metamodels, and to evaluate their performance.

2. Methodology

Metamodels are often used in engineering problems to approximate computationally intensive processes and to provide better understanding of relationships between predictors and the response. Wang and Shan [1], and Simpson et al. [2] give an overview of the most common metamodeling strategies and methods. First, a simulation experiment based on the complex model was performed to produce the data for metamodeling. This experiment is described hereunder. At this stage, a full factorial experimental design was used to better distinguish the effect of each factor. Various metamodel types exist, differing in fitting algorithms, simplicity and flexibility, with performance depending on the problem. Polynomial regression has been extensively studied and used in many applications, as it is easy to interpret and implement in any statistical software. Such model has the form $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$, $\varepsilon \sim iid\ \mathcal{N}(0, \sigma^2)$, where y is the response variable, x_1 to x_k are the explanatory variables or predictors, β_0 is the intercept, β_1 to β_k the regression coefficients corresponding to the predictors, and ε is the random error term. The model is linear in the coefficients β, but predictors may include interactions and quadratic or higher degree polynomial expressions of the variables x. Logistic regression is a generalized linear model, using the same basic formula and the logit link function to model the probability of a categorical outcome. Both models were employed in this paper. To test model performance and select among variants, 10-fold cross-validation was used.

Two indicators have been selected for metamodeling, namely the total annual feeder electricity demand (E_D), and the feeder absolute minimum voltage (U_{min}). The first is a measure of the average expected load for feeder sizing. U_{min} can be used to detect feeders with potential overloading due to heat pumps and back-up electric elements. Standard EN 50160 [3] prescribes lower voltage limits at 0.85 pu at all times, and 0.9 pu during 95 % of time each week.

2.1. Simulation experiment

The entire simulation framework for analysis of grid impact in residential neighborhoods has been developed previously. All information on models and assumptions can be found in Refs. [4] and [5]. The framework provides the models and experimental design to simulate residential distribution grids for a variety of household loads and generation, and evaluate grid performance indicators. All simulations of buildings with heat pumps, the PV generation and the network are carried out in Dymola, using the Modelica `IDEAS` library, while stochastic occupant behavior is included from the StROBe package of openIDEAS [6]. This approach allows for detailed models of thermal systems, capturing their dynamic behavior in high resolution to provide input for the electrical simulations. One-year simulations are carried out for typical Belgian weather conditions.

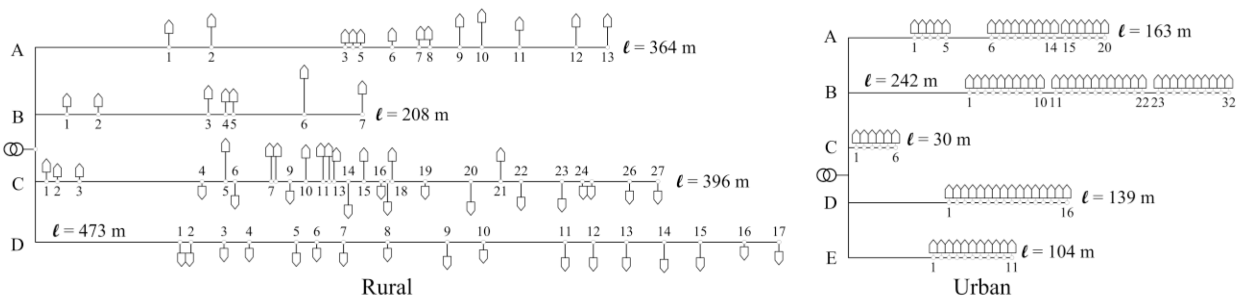

Fig. 1. Simulated rural (left) and urban (right) distribution islands with 4 and 5 feeders respectively, representing typical Belgian grids [7].

Table 1. Factors and their levels, for full-factorial design.

Factor	*T*	*Q*	*Ca*	*HP*	*PV*	T_{kVA}	U_{ref}
Description	Neighbor-hood type	Insulation quality	Cable strength	Heat pump penetration rate (%)	PV penetration rate (%)	Transformer rated capacity (kVA)	Reference transf. voltage, (pu [d])
Levels	rural, urban	new, renov., old [a]	weak, strong [b]	0, 20, 40, 60 [c]	0, 40, 60, 80	160, 250, 400	0.95, 1, 1.05

[a] Different sampling of buildings, with resulting mean *U*-value at feeder level (on average): new: 0.25, renovated: 0.38, old: 0.55 in W/(m^2K).
[b] Strong: 4×Al 150 mm^2; weak: buildings<15: 4×Al 70 mm^2, buildings≥15: 4×Al 120 mm^2.
[c] For metamodeling E_D, additional levels at 10, 30 and 50 % were included.
[d] pu: per-unit system: voltage as fraction of nominal voltage U_n = 230 V.

In particular, 300 buildings were generated, based on sampling of their geometric and thermal properties from predefined distributions. They were then simulated in Dymola as two-zone detailed building models with air-source heat pump. Each building is assigned a set of occupant profiles and is optionally equipped with a rooftop PV system. The latter consists of pre-simulated generation profiles, adjusted and scaled to the orientation and size of each system. Building simulations are performed independently of the electrical grid, and form a set of load and generation profiles.

The considered low-voltage (LV) grids, in this paper, consist of two *typical* Belgian distribution islands with 4 and 5 feeders respectively (Fig. 1). They are simulated for varying degrees of heat pump and PV penetration, different network properties and buildings of diverse thermal requirements. The full factorial design simulates all combinations of factor levels in Table 1, the latter having been derived for Belgian LV grids. For each resulting case, buildings with heat pumps and optional PV are sampled from the previously generated set. Furthermore, building sampling is repeated 10 times, to obtain a building-related uncertainty band around the expected average response [5]. The grids are simulated at 5-min resolution, as three-phase networks with unbalanced loading, using a quasi-stationary method.

3. Metamodeling electricity demand E_D

The total feeder electricity demand depends on the number of consumers and their individual consumption patterns. It consists of a random part representing the typical household loads for appliances and lighting, and a part for heating via heat pumps with back-up electric elements. Potential electricity produced by PV or other generators was not subtracted from this calculation. It is natural to expect electricity demand to increase approximately linearly with the amount of consumers and presence of heat pumps. Therefore, multiple linear regression was applied, with as candidate explanatory variables the heat pump penetration rate *HP*, neighborhood type *T*, and neighborhood insulation quality *Q* of Table 1, as well as the number of buildings in the feeder *N* (see Fig. 1). Graphical inspection of the relationships with E_D also suggest a polynomial fit. The model predicts the mean expected demand, with random error representing the uncertainty due to specific building characteristics and occupants, usually unknown in large-scale studies. Because preliminary results revealed significant misprediction of new data with *HP* between the levels of Table 1, additional levels at 10, 30 and 50 % *HP* were included in the dataset for E_D only. Over the entire simulated dataset, 99 % of E_D is included within 13.4 to 149.0 MWh/y, with mean of 57.7 and median of 51.5 MWh/y. Even though the total distribution is moderately skewed, per feeder and for constant *HP* and *Q* the distribution is close to normal.

Building a regression model is an iterative process, requiring model adequacy checking and validation, both of which may lead to re-specification of the model [8]. We describe this procedure for the metamodel of E_D hereunder. Initially, a model with only linear terms was investigated and compared with one including interactions and quadratic terms. The latter was defined using stepwise selection with minimization of the Bayesian Information Criterion (BIC).

Table 2. Summary of models and performance metrics. Root mean square (percentage) errors RMS(P)E are given relative to the original scale.

Model	**Formula** (Wilkinson notation [9])	**Train RMS(P)E**		**Test RMS(P)E**		**Adj. R^2** [a]
		MWh/y	%	MWh/y	%	
Linear	$E_D \sim 1 + T + N + HP + Q$	8.49	21.3	8.68	25.6	0.928
Quadratic	$E_D \sim 1 + T*N + T*HP + T*Q + N*HP + N*Q + HP*Q + HP^2$	5.29	11.4	5.29	11.4	0.972
Weighted	$E_D \sim 1 + T*N + T*HP + T*Q + N*HP + N*Q + HP*Q$	5.43	10.6	5.37	10.7	0.976
Log-linear	$\ln E_D \sim 1 + T*N + T*HP + T*Q + HP*Q + N^2 + HP^2$	5.15	10.4	5.45	10.4	0.970
Box-Cox	$(E_D^\lambda - 1)/\lambda \sim 1 + T*N + T*HP + T*Q + N*HP + N*Q + HP*Q + N^2 + HP^2$	5.08	10.2	5.13	10.0	0.973

[a] Adjusted R^2 is only comparable between two nested models, here the *linear* and *quadratic*. For other models, it refers to the transformed scale.

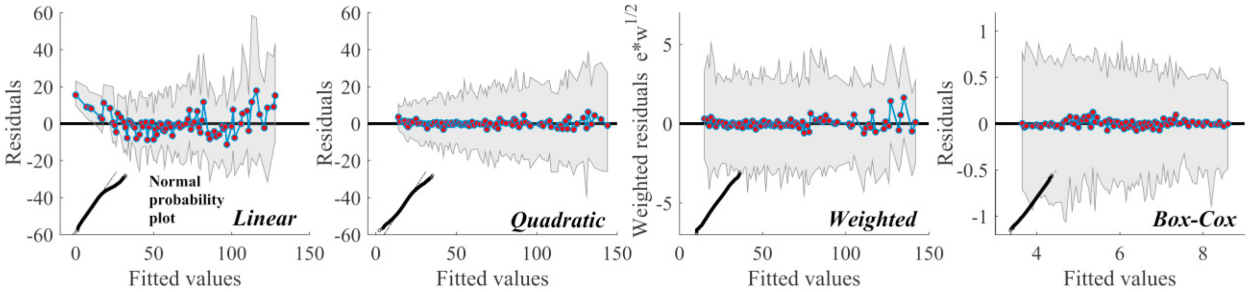

Fig. 2. Residual vs fitted response plots for the indicated models. The mean, max and min residuals per group of similar predicted E_D.

For the two model forms, 10-fold cross-validations was used for model fitting. The dataset, comprising 155 520 observations, was partitioned in 10 equally sized subsets, all containing one building sampling repetition per end-case (see section 2.1). Models were trained 10 times, each time leaving one subset out. A test dataset of 1 350 observations was additionally generated with random values for *N* and *HP*, to evaluate predictive performance. Figures are based on results of one iteration, while reported metrics for comparison are averaged over all 10 iterations, or the test set.

The equation for the *linear* model is $E_D = -14.1 - 6.3T_U + 3.5N - 0.5HP + 6.6Q_{ren} + 6.8Q_{old} + \varepsilon$. All coefficients were significant (*p*-value=0) and had standard errors below 1 %. The presented coefficients are averaged over the 10 cross-validation folds, for which the range between folds was smaller than 1 % of the mean. We only interpret coefficients of the *linear* model as example, as it is the simplest. Coefficients for categorical variables (*T*, *Q*) define the expected change in E_D, when switching from the reference to a different category. For example, urban feeders have on average 6.3 MWh/y less electricity demand than rural ones, because they have a majority of more compact houses with lower heating needs. Coefficients of continuous variables (*N*, *HP*) indicate their unique effect on E_D per unit change. For instance, an additional building would yield extra 3.5 MWh/y demand on average. Interaction terms define different effect of one variable on the response, depending on the category or value of another, therefore allowing modeling of more complex relationships.

As shown in Table 2, both *linear* and *quadratic* models have high adjusted R^2 values, 93 % and 97 % respectively, signifying they explain most of the variance of E_D. The *quadratic* model, containing interactions, performs better both in terms of explained variance and training and test errors (RMSE, RMSPE). Even the better model, however, is on average 11.4 % off predicting the response. Fig. 2 presents residual plots for each model. The *linear* model residuals show a curvilinear pattern, also indicating a missing higher-order term. Additionally, two important assumptions of linear regression are violated, namely that of constant variance and normality of the residuals. This would suggest that either model is inadequate and inefficient, possibly yielding inaccurate predictions and confidence intervals. Two methods are commonly used to correct error variance, namely, variable transformation and weighted least squares [8].

Weighted least-squares regression can reduce heteroscedasticity once the structure of variance is known. Since building sampling was repeated 10 times per combination of factors (see section 2.1), the variance of E_D could be calculated for each group. It was found to increase approximately linearly with E_D, and its reciprocal was therefore used as weight on the observations of each group. Response transformations can be used for variance stabilization as well. However, they render interpretation of the model more difficult. Additionally, predictions are in the scale of the transformed variables, and direct inverse transformation gives an estimate of the median of the response distribution instead of the mean. Solutions to estimate the mean can be found in literature [8], but were not applied in the case of E_D, as the median will be very close to the mean, since E_D is approximately normally distributed for fixed values of the predictors. Both Box-Cox transformation with $\lambda = 0.2$ and logarithmic ($\lambda = 0$) were evaluated, the former as optimal for normality of E_D, and the latter as more common alternative and easier to interpret.

Table 2 gives the RMS(P)E, expressed in the original scale, for the weighted and transformed models. These are similar with the *quadratic* model RMS(P)E, and all three new models explain the transformed variables well enough (adj. R^2 in respective scales). Major improvement can be seen on the model residuals in Fig. 2, where both the *weighted* and *Box-Cox* models gave approximately homoscedastic normal residuals in the model scale. At the same time, prediction intervals for the test set are improved, as displayed in Fig. 3a for the *quadratic* and *Box-Cox* models. For

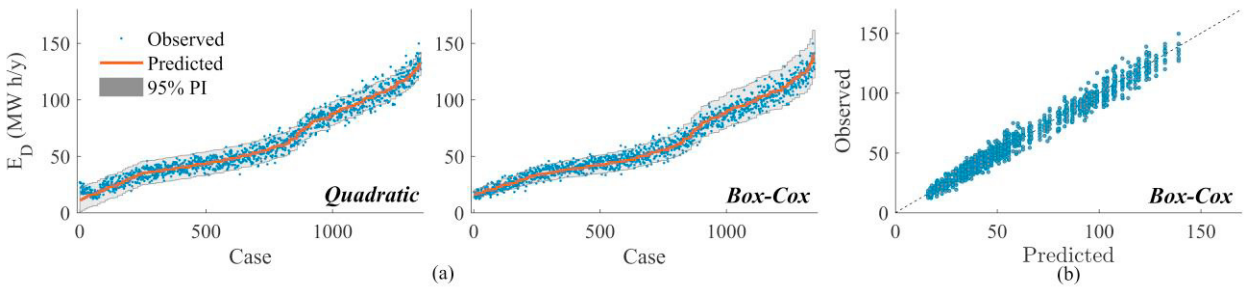

Fig. 3. For the test set: (a) observed and predicted E_D with 95 % prediction intervals, ordered by predicted value. (b) Predicted vs observed.

the *log-linear* model, heteroscedasticity was still present, rendering the model also not appropriate. Selection between the *weighted* or *Box-Cox* model is difficult, as they have different advantages and requirements, as mentioned earlier, but similar performance. Fig. 3b shows, as example, the fit of the *Box-Cox* model to the data. Given the low input detail, and relatively simple implementation, the model performs quite well in predicting the *median* expected demand per feeder. For higher accuracy, further information regarding occupancy and building properties would be necessary.

4. Metamodeling grid minimum voltage U_{min} and voltage violations

Fig. 4a shows the distribution of U_{min} is strongly left skewed, with only few violations of the lower voltage limits. In a same approach as for E_D, it was attempted to develop a linear regression metamodel for the value of absolute minimum feeder voltage U_{min}. For this model, all factors of Table 1 were used in addition to the number of buildings *N*. Furthermore, the cable cross-sectional area for each individual feeder was considered as continuous variable instead of the categories *weak* or *strong* (see footnotes in Table 1). The linear regression model with this set of predictors was found insufficient, because U_{min} importantly depends on the system dynamics and random effects such as the location and simultaneity of loads in the grid. Response transformations did not sufficiently improve the residuals' heteroscedasticity and non-normality. The weighted regression performed better in terms of (weighted) residuals. However, residuals in the original scale were biased and often large, and the fit was poor in the area of interest, below 195.5 V (Fig. 4b). In this case, different methods should be examined, for instance focusing on distribution tails.

For this paper, a classification approach was taken instead. The continuous response U_{min} was dichotomized into cases violating the limit or not: $U'_{min} = 1$ for $U_{min} < 0.85\ pu$ and $U'_{min} = 0$ otherwise. Binomial logistic regression was then used to predict the probability of a feeder having low voltage violation. Stepwise model selection with 10-fold cross-validation was used. For this problem, the original dataset with 77 760 observations was used, and test errors were calculated for the validation subsets of the cross-validation folds only. Because of the high class imbalance (3:100), we additionally compute models gradually excluding the lower levels of *HP* where violations are rare, and compare the performance. Since the logistic model does not make distributional assumptions, we assess the fit based on the performance on the validation set. This is evaluated by means of recall (correctly predicted violations over all

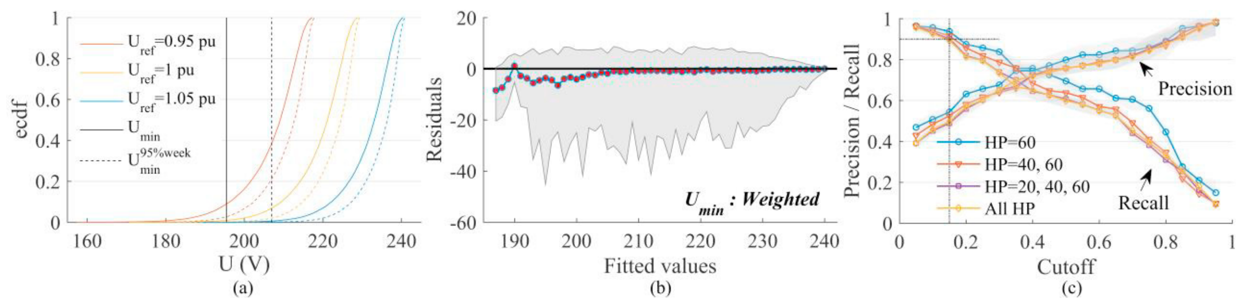

Fig. 4. (a) Empirical cumulative distribution function of minimum voltage U_{min} and minimum of 5th weekly percentiles $U_{min}^{95\%week}$, for different levels of U_{ref}. (b) Residual at original scale vs fitted response plot for the weighted model. The mean, max and min residuals per group of similar predicted U_{min}. (c) Precision and recall vs cutoff, per data subset. The mean and range among cross-validation folds are indicated.

violations) and precision (correctly predicted violations over predicted violations). The accuracy (correct predictions over total cases) does not provide useful conclusions in cases with such large class imbalance.

Based on the stepwise selection, models of the following form were fitted to the different data subsets: $logit(U'_{min}){\sim}1 + U_{ref} * Ca + T_{kVA} * Ca + HP * N + T_{kVA} * T + N * T + T_{kVA}^2 + N^2$. Fig. 4c shows the precision and recall based on the validation sets, for different cutoff values used to classify the resulting probabilities as 1 or 0. To detect the most feeders with violations, one would maximize the recall, and to avoid false detections, high precision is needed. Depending on the objectives and related costs, different cutoff values may be chosen. For both metrics, the model trained on data with only *HP*=60 % gave the best results, as it contained data with higher proportion of feeders with violation (9:100). Similar improvement could be obtained by limiting the dataset to cases with reference voltage U_{ref}=0.95 pu. Based on the best model in Fig. 4c, in order to detect more than 90 % of feeders with violation, a cutoff of 0.15 is necessary. This would also mean that 45 % of cases identified as positive (with violation) were false positives (precision 55 %). Nevertheless, all false positive cases had U_{min} between 0.85 and 0.91 pu, also very close to the limit, and potentially requiring attention. In general, the logistic model's performance can be considered acceptable, and would be sufficient to provide a first detection of feeders with risk of voltage problems under the studied conditions.

5. Conclusion

This paper aimed to analyze the potential of metamodels in approximating detailed simulation results for electricity demand and grid restrictions in residential neighborhoods. Previously developed district simulation models were used to generate the datasets for metamodeling. Simple statistical models were then developed and evaluated. Multiple linear regression for annual electricity demand requires particular attention, as basic assumptions were violated. This analysis emphasizes the importance of model adequacy checking, as opposed to simple reporting of goodness-of-fit metrics, such as R^2. In this case, weighted regression and variable transformations proved to improve residual distribution problems. Similar solutions could not be used, however, for the minimum voltage, due to its highly skewed distribution. As an alternative, the probability of voltage violation was modeled using logistic regression. This model is easy to construct, and performed well in detecting feeders with voltage problems. The developed metamodels utilized only basic metamodeling techniques available to approximate detailed simulation results for large-scale studies. Additionally, their validity is limited to the specific boundary conditions of the simulation experiment. However, they provide a baseline for metamodeling attempts in the field of building-related grid impact analysis.

Acknowledgements

This work has been conducted within the EFRO-SALK project, which receives the support of the European Union, the European Regional Development Fund, Flanders Innovation & Entrepreneurship and the Province of Limburg.

References

[1] G. Wang, S. Shan, Review of metamodeling techniques in support of engineering design optimization, *J. Mech. Des.*, 2007; 129(4):370–380.
[2] T. W. Simpson, J. D. Peplinski, P. N. Koch, and J. K. Allen, Metamodels for computer-based engineering design: survey and recommendations, *Eng. Comput.*, 2001; 17(2):129–150.
[3] EN 50160, Voltage characteristics of electricity supplied by public distribution systems, 2000.
[4] C. Protopapadaki and D. Saelens, Heat pump and PV impact on residential low-voltage distribution grids as a function of building and district properties, *Appl. Energy*, 2017; 192C:268–281.
[5] C. Protopapadaki and D. Saelens, Sensitivity of low-voltage grid impact indicators to modeling assumptions and boundary conditions in residential district energy modeling, Accepted for: *BS2017, the 15th International Conference of IBPSA*, 2017.
[6] R. Baetens, R. De Coninck, F. Jorissen, D. Picard, L. Helsen, and D. Saelens, OpenIDEAS – an open framework for integrated district energy simulations, in Proceedings of: *BS2015, the 14th International Conference of IBPSA*, 2015, p. 347–354.
[7] R. Baetens, On externalities of heat pump-based low-energy dwellings at the low-voltage distribution grid, KU Leuven, 2015.
[8] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*, 5th ed. New Jersey: John Wiley & Sons; 2012.
[9] G. N. Wilkinson and C. E. Rogers, Symbolic Description of Factorial Models for Analysis of Variance, *Applied Stat.*; 1973; 22(3):392–399.